3.4.1 The Doubly Robust Estimator

It turns out there exists a bias-corrected estimator, whose validity is based on randomization, yet which can incorporate regression predictions to increase efficiency:

$$\widehat{\psi} = \mathbb{P}_n \left[\left\{ \widehat{\mu}_1(X) - \widehat{\mu}_0(X) \right\} + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi} \right) \left\{ Y - \widehat{\mu}_A(X) \right\} \right] \tag{3.3}$$

where $\widehat{\mu}_a(x)$ is an estimate of the regression function $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$ and $\pi = \mathbb{P}(A = 1)$ is the (known) randomization probability.

The estimator (3.3) can be viewed as the plug-in estimator $\mathbb{P}_n(\widehat{\mu}_1 - \widehat{\mu}_0)$ plus a "correction" term that incorporates the randomization probabilities π. It goes by various names, including:

- model-assisted Horvitz-Thompson,
- bias-corrected plug-in,
- semiparametric or semiparametric efficient,
- augmented inverse-probability-weighted (AIPW),
- doubly robust.

We will see variants of the estimator (3.3) throughout the course, and will mostly refer to it as doubly robust. It has an interesting and somewhat difficult-to-trace history across subfields of statistics. Here is an abbreviated and limited portion of its path across the literature:

- In survey sampling problems, Cochran [1977] and others used simple regression models in an agnostic way to improve the efficiency of the unbiased Horvitz-Thompson estimator from 1952.
- Robins and Rotnitzky [1995], Robins et al. [1994, 1995] studied efficient semiparametric estimation in general missing data problems (extending work by Bickel et al. [1993] and Pfanzagl [1982] and others), and presented a version of this estimator (3.3) where nuisance quantities were estimated with parametric models.
- Robins and Wang [2000] started referring to the estimator (3.3) as "doubly protected", and Robins and Rotnitzky [2001] and Bang and Robins [2005] as "doubly robust".
- In a series of papers, Tsiatis and colleagues [Davidian et al., 2005, Leon et al., 2003, Yang and Tsiatis, 2001, Zhang et al., 2008] applied the theory from Robins and others to randomized experiments, focusing on efficiency concerns. These papers are a nice introduction to the estimator (3.3) in the experimental setup.
- In the early to mid 2000s, van der Laan and Robins [2003] and others started developing theory for the case where nuisance estimators such as $\widehat{\mu}_a$ are estimated nonparametrically.

- The estimator and related methods have been recently re-discovered in the econometrics world [Chernozhukov et al., 2018], with more of a focus on high-dimensional sparse models.

In fact it can be shown that any (regular) $\sqrt{n}$-consistent and asymptotically normal estimator can be written in the form (3.3), for some choice of $\widehat{\mu}_a$. So in fact we have already seen some variants of it, e.g.:

- The difference-in-means estimator is recovered if $\widehat{\mu}_a = \mathbb{P}_n(Y \mid A = a)$, and
- the Horvitz-Thompson or inverse-probability-weighted estimator if $\widehat{\mu}_a = 0$.

In fact, shortly we will study some cases where, surprisingly, the parametric plug-in takes this form with for example $\widehat{\mu}_a = g(\widehat{\beta}_0 + \widehat{\beta}_1 a + \widehat{\beta}_2^{\mathrm{T}} x)$. This is one of the reasons it is a bit unclear where the estimator originated, since it includes many variants as a special case.

Remark 3.4. As we did above, at several points in this section we will refer to *regular* estimators. A more detailed discussion will come later, but for the time being a regular estimator can be taken to mean an estimator whose limiting distribution is insensitive to local perturbations of the data-generating process. Imposing regularity rules out *super-efficient* estimators, for example, which trade very good performance at a particular $\mathbb{P}$ for very bad performance "near" $\mathbb{P}$. More discussion can be found in Tsiatis [2006] and van der Vaart [2000].

As mentioned earlier, the estimator (3.3) can be interpreted as a corrected version of the plug-in estimator $\widehat{\psi}_{pi} = \mathbb{P}_n(\widehat{\mu}_1 - \widehat{\mu}_0)$ since

$$\widehat{\psi} = \widehat{\psi}_{pi} + \mathbb{P}_n \left[\left(\frac{A}{\pi} - \frac{1-A}{1-\pi} \right) \left\{ Y - \widehat{\mu}_A(X) \right\} \right].$$

We will see how the correction term removes any bias afflicting the regression estimator $\widehat{\mu}_a$. The doubly robust estimator can also be viewed as a corrected (of "augmented") version of the inverse-probability weighted (Horvitz-Thompson) estimator $\widehat{\psi}_{ipw} = \mathbb{P}_n\{\left(\frac{AY}{\pi} - \frac{1-A}{1-\pi}\right) Y\}$ since

$$\widehat{\psi} = \widehat{\psi}_{ipw} + \mathbb{P}_n \left[\left(1 - \frac{A}{\pi} \right) \widehat{\mu}_1(X) - \left(1 - \frac{1-A}{1-\pi} \right) \widehat{\mu}_0(X) \right].$$

We know from the previous chapter that $\widehat{\psi}_{ipw}$ is already unbiased; thus the above augmentation term is reducing variance rather than bias.

Here is example code showing how to correct the plug-in estimator we constructed earlier:

```
> cbind(x,a,y)[1:5,]
                x a y
 [1,] -0.44577826 0 0
 [2,] -1.20585657 0 0
 [3,]  0.04112631 1 1
 [4,]  0.63938841 0 0
 [5,] -0.78655436 0 1
>
> mumod <- glm(y~x+a, family=binomial)
> mu1hat <- predict(mumod, newdata=data.frame(x,a=1) ,type="response")
> mu0hat <- predict(mumod, newdata=data.frame(x,a=0), type="response")
>
> pi <- 0.5; muahat <- a*mu1hat + (1-a)*mu0hat
>
> mean( (mu1hat-mu0hat) + (a/pi - (1-a)/(1-pi)) * (y-muahat) )
[1] 0.2303284
```

Remark 3.5. Note that the doubly robust estimator requires no extra model fitting beyond that already required to construct the plug-in estimator.

A natural question about the doubly robust estimator is: where does the correction come from, and why does it take that specific form? A complete answer to this is highly non-trivial; we will pursue it in depth in later chapters. However some short discussion is still useful. The form of the correction comes from nonparametric efficiency theory for functional estimation [Bickel et al., 1993, Tsiatis, 2006, van der Laan and Robins, 2003], and there are two high-level heuristics for thinking about it. The first is that the average treatment effect parameter $\psi = \psi(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}(\mu_1 - \mu_0)$ is a "smooth" functional, when viewed as a map from probability distributions $\mathbb{P}$ to the real line; and this smoothness allows for convenient and effective bias correction. The second is that a randomized experiment with known treatment mechanism leads to a semiparametric model for the distribution $\mathbb{P}$ from which we sample: part of the distribution $\mathbb{P}$ is known (the conditional distribution of treatment given any covariates) while the rest is left unrestricted (the covariate distribution and the conditional distribution of the outcome given covariates and treatment). Under this semiparametric model, one can use tools from efficiency theory to derive the form of *all* possible (regular) asymptotically normal estimators of the parameter ψ, and subsequently find the one with the smallest variance.

Answering the question of *why* the correction works is easier than answering where it comes from. This is the focus of the next section.

3.4.2 Properties of the Doubly Robust Estimator

Here we study the bias, variance, and limiting distribution of the estimator (3.3).

Remark 3.6. In this section we are going to consider the case where the regression estimator $\widehat{\mu}_a$ is constructed from a separate training sample D^n independent of the experimental sample $Z^n = \{(X_1, A_1, Y_1), ..., (X_n, A_n, Y_n)\}$. This setup can be accomplished easily in practice by simply randomly splitting the sample, and using half as D^n for training and the other half as Z^n for estimation. Note that in this case, variance results should really be framed in terms of $n/2$ instead of n; if this loss of efficiency is concerning to you, luckily there is an easy fix: after constructing the sample-split estimator, swap the samples, using Z^n for training and D^n for estimation, and then average the resulting estimators. This approach will recover full sample size efficiency.

There are two reasons for doing sample splitting: the first is that the analysis is more straightforward, and the second more important reason is that it prevents overfitting and allows for the use of arbitrarily complex estimators $\widehat{\mu}_a$ (e.g., random forests, boosting, neural nets). Without sample splitting, one would have to restrict the complexity of the estimator $\widehat{\mu}_a$ via empirical process conditions (e.g., via Donsker class or entropy restrictions). Intuitively, this is because the estimator $\widehat{\psi}$ is using the data twice: once to estimate the unknown function μ_a and once to estimate the bias correction. Sample splitting ensures that these tasks are accomplished independently.

As in our analysis of the plug-in estimator in the previous section, we note that our estimator can be written as a sample average of an estimated function. Namely $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$ where now $\widehat{f} = f(\widehat{\mu}) \equiv f_1(\widehat{\mu}) - f_1(\widehat{\mu})$ for

$$f_a(\overline{\mu}) \equiv \overline{\mu}_a(X) + \frac{\mathbb{1}(A=a)}{\mathbb{P}(A=a)}\Big\{Y - \overline{\mu}_A(X)\Big\} \tag{3.4}$$

First we tackle the bias of $\widehat{\psi} = \mathbb{P}_n(\widehat{f})$, *under no modeling assumptions.*

Theorem 3.2. *Consider an iid Bernoulli experiment with $\mathbb{P}(A=1) = \pi$. Then the doubly robust estimator $\widehat{\psi}$ in (3.3) is unbiased for the average treatment effect when the regression estimates $\widehat{\mu}_a$ are constructed from a separate independent sample.*

Proof. We will derive the bias for $\psi_1 = \mathbb{E}(Y^1)$ with $\widehat{\psi}_1 = \mathbb{P}_n(\widehat{f}_1)$ since the logic is exactly the same for $\mathbb{E}(Y^0)$ and the difference $\psi = \psi_1 - \psi_0$. First note that for any $\overline{\mu}_1$

$$\begin{aligned}
\mathbb{P}\{f_1(\overline{\mu})\} &= \mathbb{P}\left[\overline{\mu}_1(X) + \frac{A}{\pi}\Big\{Y - \overline{\mu}_1(X)\Big\}\right] \\
&= \mathbb{P}\left[\overline{\mu}_1(X) + \frac{\pi}{\pi}\Big\{\mu_1(X) - \overline{\mu}_1(X)\Big\}\right] \\
&= \mathbb{E}\{\mu_1(X)\} = \psi_1
\end{aligned} \tag{3.5}$$

where the second equality used iterated expectation and the Bernoulli randomization. Therefore we have

$$\mathbb{E}(\widehat{\psi}_1 \mid D^n) = \mathbb{P}\{f(\widehat{\mu}_1)\} = \psi_1$$

where the first equality uses the fact that $\widehat{\mu}_a(x)$ is fixed given independent D^n and the iid assumption, and the second (3.5). □

Theorem 3.2 is a simple but powerful result. It shows the doubly robust estimator is exactly unbiased, *for any choice of* regression estimator $\widehat{\mu}_a$. Hence, although the estimator $\widehat{\psi}$ exploits covariate information, its bias is not at all affected by accidentally misspecified models or biased regression estimators with slow convergence rates.

Remark 3.7. Theorem 3.2 also has an important implication for understanding the variance and limiting distribution of $\widehat{\psi}$. Namely, the logic in the proof shows that

$$\mathbb{P}\{f(\overline{\mu})\} = \psi$$

for *any* (fixed) $\overline{\mu}$. This means that, since $\widehat{\psi}$ is a sample average of an estimated function and thus the decomposition from Lemma 3.1 holds, we can write

$$\begin{aligned} \widehat{\psi} - \psi &= (\mathbb{P}_n - \mathbb{P})(\widehat{f} - \overline{f}) + \mathbb{P}(\widehat{f} - \overline{f}) + (\mathbb{P}_n - \mathbb{P})\overline{f} \\ &\equiv T_1 + T_2 + Z^* \end{aligned} \tag{3.6}$$

for *any* $\overline{f} = f(\overline{\mu})$. Since it will be useful in our analysis for $\widehat{f}$ to be consistent for $\overline{f}$, we will simply define $\overline{f} = f(\overline{\mu})$ to be the corresponding probability limit, i.e., by taking $\overline{\mu}_a$ to be a fixed function such that $\|\widehat{\mu}_a - \overline{\mu}_a\| = o_{\mathbb{P}}(1)$. We will see that this will allow us to completely sidestep whether the estimator $\widehat{\mu}_a$ is consistent for the *true* regression function μ_a, and instead just require that it be consistent for *something*.

Now we tackle the limiting distribution of $\widehat{\psi}$. Recall we know Z^* in the decomposition (3.6) is asymptotically normal, so we only need to understand the T_1 and T_2 terms. First we provide a general analysis of the first term

$$T_1 = (\mathbb{P}_n - \mathbb{P})(\widehat{f} - \overline{f})$$

in that decomposition.

Lemma 3.3. *Let $\mathbb{P}_n$ denote the empirical measure over $Z^n = (Z_1, ..., Z_n)$, and let $\widehat{f}(z)$ be any function estimated from a sample $D^N = (Z_{n+1}, ..., Z_{n+N})$, which is independent of Z^n. Then*

$$(\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - f\|}{\sqrt{n}}\right).$$

Proof. See Kennedy et al. [2019a]. □

Lemma 3.3 shows that T_1 terms are asymptotically negligible, i.e., that $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$, as long as $\widehat{f}$ is consistent for f (or $\overline{f}$ in our case, which will hold by definition).

The next result gives the limiting distribution of the doubly robust estimator, under no assumptions beyond the experiment design (and iid sampling) and that the regression estimators $\widehat{\mu}_a$ converge to anything at any rate.

Theorem 3.3. *Consider an iid Bernoulli experiment with $\mathbb{P}(A=1)=\pi$. Suppose the regression estimators $\widehat{\mu}_a$ are:*

1. *constructed from a separate independent sample, and*
2. *consistent (at any rate) for some functions $\overline{\mu}_a$ (not necessarily the true regression functions μ_a) in the sense that $\|\widehat{\mu}_a - \overline{\mu}_a\| = o_{\mathbb{P}}(1)$.*

Then the doubly robust estimator $\widehat{\psi}$ is root-n consistent and asymptotically normal with

$$\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N\Big(0, var(\overline{f})\Big)$$

where $\overline{f} = f(\overline{\mu})$ is defined as in (3.4).

Proof. By Lemma 3.1 we can write the decomposition (3.6) with $\overline{f} = f(\overline{\mu})$ for any $\overline{\mu}$. We will define $\overline{\mu}$ as the probability limit of $\widehat{\mu}$, as in the statement of the theorem.

By Lemma 3.3, we have $T_1 = O_{\mathbb{P}}(\|\widehat{f} - \overline{f}\|/\sqrt{n})$. Now note

$$\begin{aligned}
\|\widehat{f}_1 - \overline{f}_1\|^2 &= \left\| \widehat{\mu}_1 + \frac{A}{\pi}\Big\{Y - \widehat{\mu}_1(X)\Big\} - \overline{\mu}_1 - \frac{A}{\pi}\Big\{Y - \overline{\mu}_1(X)\Big\} \right\|^2 \\
&= \left\| \Big\{\widehat{\mu}_1 - \overline{\mu}_1\Big\} \left\{1 - \frac{A}{\pi}\right\} \right\|^2 \\
&= \int \Big\{\widehat{\mu}_1(x) - \overline{\mu}_1(x)\Big\}^2 \left(\frac{A-\pi}{\pi}\right)^2 \, d\mathbb{P}(z) \\
&= \left(\frac{\text{var}(A)}{\pi^2}\right) \int \Big\{\widehat{\mu}_1(x) - \overline{\mu}_1(x)\Big\}^2 \, d\mathbb{P}(x) \\
&= \left(\frac{1-\pi}{\pi}\right) \|\widehat{\mu}_1 - \overline{\mu}_1\|^2
\end{aligned}$$

where the fourth equality used the Bernoulli randomization. The same logic applies to $\|\widehat{f}_0 - \overline{f}_0\|$, and so by the triangle inequality

$$T_1 = O_{\mathbb{P}}\left(\frac{\|\widehat{f} - \overline{f}\|}{\sqrt{n}}\right) = O_{\mathbb{P}}\left(\frac{\|\widehat{\mu}_1 - \overline{\mu}_1\| + \|\widehat{\mu}_0 - \overline{\mu}_0\|}{\sqrt{n}}\right)$$

which is $o_{\mathbb{P}}(1/\sqrt{n})$ since $\|\widehat{\mu}_a - \overline{\mu}_a\| = o_{\mathbb{P}}(1)$ by definition.

For the T_2 term, we have $\mathbb{P}(\widehat{f} - \overline{f}) = 0$ by (3.5). This gives the result. □

Theorem 3.3 shows that not only is the doubly robust estimator $\widehat{\psi}$ unbiased for any choice of regression estimator, it is also root-n consistent and asymptotically normal – even if the estimators $\widehat{\mu}_a$ are completely misspecified, and or converging at arbitrarily slow rates. This is a pretty amazing result.

This immediately implies that distribution-free confidence intervals can be constructed as in the following corollary.

Corollary 3.2. *Under the assumptions of Theorem 3.3, a distribution-free asymptotic 95% confidence interval for the average treatment effect ψ is given by*

$$\widehat{\psi} \pm 1.96\sqrt{\frac{\widehat{var}\{f(\widehat{\mu})\}}{n}}.$$

Further, finite-sample variance bounds can be constructed using the same logic as in the proof of Theorem 3.3.

Proposition 3.3. *Under the assumptions of Theorem 3.3, the doubly robust estimator $\widehat{\psi}$ in (3.3) has variance at most*

$$var(\widehat{\psi}) \leq \frac{1}{n}\left\{var(\overline{f}) + \left(\frac{1-\pi}{\pi}\right)\|\widehat{\mu}_1 - \overline{\mu}_1\|^2 + \left(\frac{\pi}{1-\pi}\right)\|\widehat{\mu}_0 - \overline{\mu}_0\|^2\right\}.$$

3.4.3 Efficiency

We have learned the surprising result that the sample-split doubly robust estimator is exactly unbiased for any choice of regression estimator $\widehat{\mu}_a$, and root-n consistent and asymptotically normal as long as $\widehat{\mu}_a$ converges to some fixed function at any rate. As would be expected, the efficiency of the doubly robust estimator depends on the probability limits $\overline{\mu}_a$ that the regression estimators $\widehat{\mu}_a$ converge to. This raises some important questions:

- What is the best possible (i.e., most efficient) probability limit $\overline{\mu}_a$?
- Is the doubly robust estimator necessarily more efficient than the difference-in-means or Horvitz-Thompson estimator?

Recall however that the difference-in-means and Horvitz-Thompson estimators can be written as variants of the doubly robust estimator, for particular choices of $\widehat{\mu}_a$. Therefore the best choice of $\overline{\mu}_a$ will dominate others in this class.

The next result shows what you might expect: that the best limit $\overline{\mu}_a$ in terms of efficiency is the *true* regression function μ_a (recall this limit is irrelevant for bias since $\widehat{\psi}$ is unbiased for any $\widehat{\mu}_a$).

Theorem 3.4. *Define $f(\overline{\mu})$ as in (3.4). Then for any $\overline{\mu} = (\overline{\mu}_1, \overline{\mu}_0)$ with $\overline{\mu}_a : \mathcal{X} \mapsto \mathbb{R}$*

$$var\{f(\overline{\mu})\} \geq var\{f(\mu)\}$$

where $\mu = (\mu_1, \mu_0)$ denotes the true regression functions.

Proof. We have

$$\begin{aligned}
\operatorname{var}\{f(\overline{\mu})\} &= \operatorname{var}\left[\overline{\mu}_1(X) - \overline{\mu}_0(X) + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \overline{\mu}_A(X)\right\}\right] \\
&= \operatorname{var}\Bigg\{(\mu_1 - \mu_0) + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)(Y - \mu_A) \\
&\qquad + \left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\Bigg\} \\
&= \operatorname{var}\{f(\mu)\} + \operatorname{var}\left\{\left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\} \\
&\qquad + 2\operatorname{cov}\left\{f(\mu), \left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\}
\end{aligned}$$

But the latter covariance is zero since

$$\begin{aligned}
&\operatorname{cov}\left\{f(\mu), \left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\} \\
&\qquad = \mathbb{E}\left[(\mu_1 - \mu_0 - \psi)\left\{\left(1 - \frac{A}{\pi}\right)(\overline{\mu}_1 - \mu_1) - \left(1 - \frac{1-A}{1-\pi}\right)(\overline{\mu}_0 - \mu_0)\right\}\right] \\
&\qquad = 0
\end{aligned}$$

where the second equality follows from iterated expectation since $\mathbb{E}\{f(\mu) \mid X, A\} = \mu_1 - \mu_0$, and the third since $A \perp\!\!\!\perp X$ so that $\mathbb{E}\{Ag(X)\} = \pi\mathbb{E}\{g(X)\}$ for any g. This gives the result □

Theorem 3.4 is critically informative about how to construct the doubly robust estimator $\widehat{\psi}$ in practice. Namely, it indicates that we should estimate the regression functions as flexibly as possible: bias is zero regardless, and efficiency is optimized when the regression functions are estimated consistently. This is a special case not often seen in statistics where there is essentially no penalty (at least asymptotically) for slow rates of convergence, and important benefits for consistency.

However the second question still remains: when based on a misspecified model for μ_a, does the doubly robust estimator necessarily still improve efficiency (say relative to the Horvitz-Thompson estimator)? In fact, this is not necessarily so, for particularly misspecified choices of $\widehat{\mu}_a$. However, there are multiple approaches that can be used to guarantee efficiency gains. One simple option proposed by Rubin and van der Laan [2008] is to posit a working parametric model $\mu_a(x) = \mu_a(x;\beta)$, but rather than estimating the parameters via maximum likelihood, instead estimate parameters by picking those that minimize an estimator of the variance, i.e., use

$$\widetilde{\beta} = \arg\min_{\beta} \ \widehat{\operatorname{var}}\left[\mu_1(X;\beta) - \mu_0(X;\beta) + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \mu_A(X;\beta)\right\}\right].$$

Other approaches similar in spirit are also possible [Tan, 2010].

3.5 Back to the Plug-In

In the previous section we saw strong evidence that, if one wants to remain agnostic about the data-generating process beyond the known randomization probabilities, retaining robustness while exploiting covariate information to gain efficiency, then the doubly robust estimator (3.3) using a flexible regression estimator is a good choice. In particular, it will be root-n consistent and asymptotically normal as long as the regression estimator $\widehat{\mu}_a$ converges to *anything at any rate*, and if the regression estimator $\widehat{\mu}_a$ is consistent for the true regression function (again *at any rate*) then it will be asymptotically efficient.

However, in practice, applied researchers often use ordinary least squares or plug-in estimators based on parametric models. Do we have any basis for trusting such results? In fact, the following surprising result shows that some if not many parametric plug-in estimators can be represented in the doubly robust form: they are doubly robust estimators disguised as plug-ins. (Though it is important to note that this is not true of all plug-in estimators.)

Theorem 3.5. *Suppose regression predictions $\widehat{\mu}_a$ satisfy*

$$\mathbb{P}_n\left[(1, A)^{\mathrm{T}}\left\{Y - \widehat{\mu}_A(X)\right\}\right] = 0 \tag{3.7}$$

where $A \perp\!\!\!\perp X$ is randomized according to a Bernoulli experiment. Then the parametric plug-in estimator

$$\mathbb{P}_n\left\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\right\}$$

is numerically equivalent to the doubly robust estimator

$$\mathbb{P}_n\left[\left\{\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\right\} + \left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \widehat{\mu}_A(X)\right\}\right].$$

Proof. Since

$$\mathbb{P}_n\left[(1, A)^{\mathrm{T}}\left\{Y - \widehat{\mu}_A(X)\right\}\right] = 0$$

it follows that

$$\frac{1}{\pi}\mathbb{P}_n\left[A\left\{Y - \widehat{\mu}_A(X)\right\}\right] = \frac{1}{1-\pi}\mathbb{P}_n\left[A\left\{Y - \widehat{\mu}_A(X)\right\}\right] = \frac{1}{1-\pi}\mathbb{P}_n\left[\left\{Y - \widehat{\mu}_A(X)\right\}\right] = 0.$$

Therefore

$$\mathbb{P}_n\left[\left(\frac{A}{\pi} - \frac{1-A}{1-\pi}\right)\left\{Y - \widehat{\mu}_A(X)\right\}\right] = 0$$

so that the correction term in the doubly robust zero is estimator, and the plug-in and doubly robust estimator are equal. □

The sufficient condition (3.7) in Theorem 3.5 says that the $\widehat{\mu}_a$ residuals must average to zero both in the whole sample and among the treated. This will hold for example in generalized linear models with an intercept and main effect term for treatment. Thus Theorem 3.5 shows that, although our earlier analysis of the parametric plug-in appeared to hinge on restrictive parametric model assumptions, this is not necessarily so – at least in Bernoulli experiments, and for plug-ins based on models with an intercept and main effect for treatment. Such parametric plug-in estimators will be root-n consistent for the average treatment effect (and asymptotically normal), even under misspecification of the regression estimator $\widehat{\mu}_a$, as long as it converges in probability to anything at any rate (a very weak condition).

Remark 3.8. Although a plug-in whose regression estimates satisfy (3.7) will take on all the advantageous robustness and efficiency properties of the doubly robust estimator, note that variance estimates must be based on the doubly robust variance as in Corollary 3.2. Otherwise (e.g., if based on the parametric model being correct as in Theorem 3.1) corresponding confidence intervals and hypothesis tests may not be valid.

Remark 3.9. The condition (3.7) holding will generally not be enough to ensure that a plug-in will be doubly robust, outside of a Bernoulli experiment where treatment is completely independent of covariates. For example, this would not be sufficient in the conditionally randomized experiment discussed next, since the randomization probabilities cannot be brought outside the average as in the proof of Theorem 3.5.

3.6 Conditional Randomization

In conditionally randomized experiments, the randomization probabilities can differ by covariate values, e.g., in a stratified Bernoulli experiment one sets

$$\mathbb{P}(A = 1 \mid X = x, Y^a) = \pi(x)$$

where the function $\pi(x)$ can vary with x (recall $\pi(x) = \pi$ in a Bernoulli experiment).

Experiments may use conditional or stratified randomization to improve efficiency (e.g., by treating more units at covariate values where treated outcomes are more variable than control outcomes), or to improve subject outcomes (e.g., by treating more units at covariate values where treated outcomes are likely to be higher than control outcomes, and treating fewer when treatment is ineffective or even harmful).

Doubly robust estimators take the same form as (3.3), but replace π with $\pi(x)$, i.e.,

$$\mathbb{P}_n \left[\left\{ \widehat{\mu}_1(X) - \widehat{\mu}_0(X) \right\} + \left\{ \frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)} \right\} \left\{ Y - \widehat{\mu}_A(X) \right\} \right]$$

and the logic of the theoretical analysis is the same as well.

There are some important differences between simple Bernoulli experiments and conditionally randomized designs, however. First, the difference-in-means estimator is no longer a valid estimator, since it is no longer the case that $A \perp\!\!\!\perp Y^a$ or $A \perp\!\!\!\perp (X, Y^a)$; instead, in a conditionally randomized experiment it only holds that $A \perp\!\!\!\perp Y^a \mid X$. Second, plug-in estimators are not in general doubly robust in conditionally randomized designs, even when they satisfy the condition (3.7); this is because the randomization probabilities cannot be brought outside the average as in the proof of Theorem 3.5.

Appendix A

Notation Guide

Y^a	Potential outcome under treatment/exposure $A = a$
$\perp\!\!\!\perp$	Statistically independent
$\xrightarrow{p}$	Convergence in probability
$\rightsquigarrow$	Convergence in distribution
$O_{\mathbb{P}}(1)$	Bounded in probability
$o_{\mathbb{P}}(1)$	Converging in probability to zero
$\mathbb{P}_n$	Sample average operator, as in $\mathbb{P}_n(\widehat{f}) = \mathbb{P}_n\{\widehat{f}(Z)\} = \frac{1}{n}\sum_{i=1}^{n}\widehat{f}(Z_i)$
$\mathbb{P}$	Conditional expectation given the sample operator, as in $\mathbb{P}(\widehat{f}) = \int \widehat{f}(z)\, d\mathbb{P}(z)$
$\|\cdot\|$	$L_2(\mathbb{P})$ norm $\|f\| = \sqrt{\mathbb{P}(f^2)}$ or Euclidean norm, depending on context
$\|\cdot\|_1$	$L_1(\mathbb{P})$ norm $\|f\|_1 = \mathbb{P}(\|f\|)$
$\|\cdot\|_\infty$	L_∞ or supremum norm $\|f\|_\infty = \sup_z \|f(z)\|$
$\mathcal{H}(s)$	Hölder class of functions with smoothness index s
$\lesssim$	Less than or equal, up to a constant multiplier

Bibliography

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press, 1993.

D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.

M. Davidian, A. A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science*, 20(3):261, 2005.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, pages 237–249, 2008.

S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, pages 29–46, 1999.

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245, 2017.

E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics (to appear)*, 2019a.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019b.

S. Leon, A. A. Tsiatis, and M. Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055, 2003.

J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

J. Pfanzagl. *Contributions to a general asymptotic statistical theory*, volume 13. Springer, 1982.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995.

J. M. Robins and A. Rotnitzky. Comments on: Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:920–936, 2001.

J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87(1): 113–124, 2000.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

D. B. Rubin and M. J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4(1), 2008.

Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer, 2003.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.

L. Yang and A. A. Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715, 2008.